CBM First-level Event Selector data management developments*

H. Hartmann[1], *J. de Cuveland*[1], *and V. Lindenstruth*[1]
[1]FIAS Frankfurt Institute for Advanced Studies, Goethe-Universität Frankfurt am Main, Germany

The First-level Event Selector (FLES) is a high performance computing cluster functioning as the central event selection system in the CBM experiment. It combines data from a large number of input links to time intervals and distributes them to the compute nodes, via a high-performance network. Simultaneously, the FLES carries out online analyses and complete event reconstruction on the data. Data rates at this point are expected to exceed 1 TByte/s.

The FLES system will consist on one hand of a scalable supercomputer with custom FPGA-based input interface cards and a fast event-building network and will be constructed largely from standard components. On the other hand special developed software allowing to process the incoming data in real-time builds up the FLES.

A small scale, highly customizable platform, the Micro-FLES cluster was installed at GSI. Eight identical compute nodes provide a total of 192 logic cores and 512 GB memory plus one head node for infrastructural services. This test system enables studies on the development of the FLES such as elaborating performant software for timeslice building.

A *timeslice* is the fundamental data structure managing access to all detector raw data of a given time interval. In addition to existing timeslice building prototype software based on InfiniBand Verbs investigations of a more high-level interface to the network hardware have been performed using MPI. For this purpose a specialized micro benchmark test suite was developed simulating the FLES timeslice building use case. Benchmark results for simultaneous data transfer on the Micro-FLES are displayed in Fig 1. When communication is established only between three nodes, MPI's performance compares to the maximum data rate of point to point communication for Infiniband Verbs (green curve) on the Infiniband-FDR network. However, the data rate decreases by 15% when all eight nodes of the Micro-Fles are participating in an any-to-any communication. Further tests on bigger compute cluster are neccessary to evaluate the achievable data rates for MPI on a big scale and are currently under investigation.

In 2014 the FLES demonstrator system was upgraded signifciantly to the Micro-FLES2. First, the Micro-FLES2 was equipped with the latest Mellanox dual Connect-IB HCAs (mlx5), in addition to the existing Mellanox dual ConnectX-3 cards (mlx4). Overall the new cards are faster than the old as shown in Fig. 2. A data rate of 6 GB/s can be achieved using only one of the four ports, already. Furthermore, the new cards feature a 16x PCIe 3.0 interface and

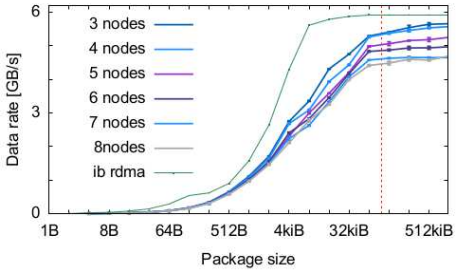

Figure 1: MPI benchmark on the Micro-FLES.

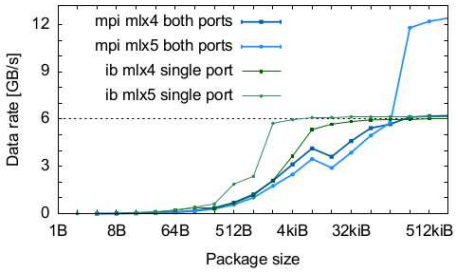

Figure 2: Perfomance tests for the Micro-FLES2.

therefore allow to saturize the full bandwith of both ports. An accumulated data rate of 18 GB/s can be achieved utilizing all Infiniband ports. With this first upgrade the Micro-FLES2 can send data from node to node three times as fast as before (e.g., 18 GB/s instead of 6GB/s). The improved performance is essential for the development of timesslice building software.

Secondly, two further Mellanox SX6036 36-port 56Gb/s switches were installed in order to realize different network setups such as a fat tree. This helps investigating routing issues in the development of software when distributing the incoming data. The previous existing switch was connected with full bidirectional bandwidth to both of the new switches making them leaves of a fat tree. All first ports of mlx4 and mlx5 for each node were connected to leaf-switch1 and all second ports to leaf-switch2. Using this setup the network structure and blocking ratio in case of a fat tree can be configured dynamically via the provided internet interface of the switches. The upgraded Micro-FLES2 provides better performance and a greater flexibilty in testing different scenarios allowing to evaluate a greater varity of possibilites for the final system – the FLES.

* Work supported by BMBF (05P12RFFCP) and HIC for FAIR